Problems and Paradoxes in High-dimensional Data

Michael Komodromos

February 7, 2022

Outline

Sources of high-dimensional data

Sources of high-dimensional data

- **Biotech**: DNA microarrays, proteomics, transcriptomics etc.

Sources of high-dimensional data

- **Biotech**: DNA microarrays, proteomics, transcriptomics etc.
- **Images / video**: medical, astrophysics, surveillance

Sources of high-dimensional data

- **Biotech**: DNA microarrays, proteomics, transcriptomics etc.
- **Images / video**: medical, astrophysics, surveillance
- **Consumer preferences**: books, movie, music recommendation

Sources of high-dimensional data

- **Biotech**: DNA microarrays, proteomics, transcriptomics etc.
- **Images / video**: medical, astrophysics, surveillance
- **Consumer preferences**: books, movie, music recommendation
- etc.

Examples

240p image $\sim 100,000$ pixels, 3 color bandwidth $p \approx 300,000$

Paradoxes

Volume of an n-D ball

- What's the volume of a circle with radius r?

Volume of an n-D ball

- What's the volume of a circle with radius r?
- πr^2

Volume of an n-D ball

- What's the volume of a circle with radius r?
- πr^2

- What's the volume of a ball with radius r?

Volume of an n-D ball

- What's the volume of a circle with radius r?
- πr^2

- What's the volume of a ball with radius r?
- $\frac{4}{3}\pi r^3$

Volume of an n-D ball

- What's the volume of a circle with radius r?
- πr^2

- What's the volume of a ball with radius r?
- $\frac{4}{3}\pi r^3$

Volume of p-dimensional ball with radius $r > 0$

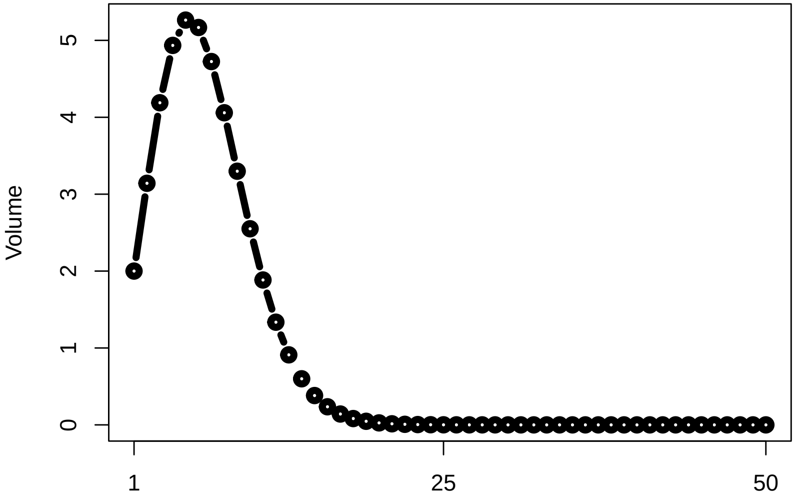
$$V(p; r) = \frac{1}{\Gamma(p/2+1)} \pi^{p/2} r^p \tag{1}$$

Does the volume get bigger or smaller as we increase p?

Or more generally, how does the volume behave?

The volume

The volume



Where is the mass concentrated?

We've noticed that the volume of a p-dimensional ball tends to 0 as p increases, then naturally we might ask:

Where is the mass concentrated?

Where is the mass concentrated?

Where is the mass concentrated?

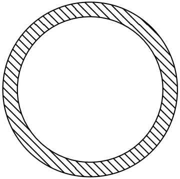
Consider the volume in the crust, i.e. the volume

$$C(p;r) = V(p;r) - V(p;0.99r) \qquad (2)$$

Where is the mass concentrated?

Consider the volume in the crust, i.e. the volume

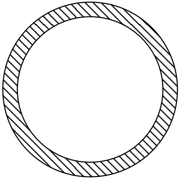
$$C(p; r) = V(p; r) - V(p; 0.99r) \qquad (2)$$



Where is the mass concentrated?

Consider the volume in the crust, i.e. the volume

$$C(p; r) = V(p; r) - V(p; 0.99r) \qquad (2)$$

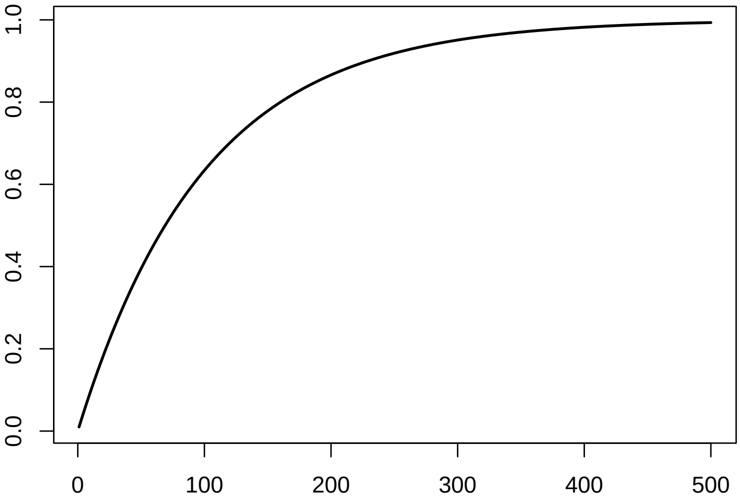


Then the volume in the crust as a fraction of the total volume, i.e.

$$\frac{C(p; r)}{V(p; r)} = 1 - 0.99^p \qquad (3)$$

Where is the mass concentrated?

Where is the mass concentrated?



So far...

So far...

Recap

- The volume of p-dimensional balls tends to 0
- The crust contains nearly all the mass

So far...

Recap

- The volume of p-dimensional balls tends to 0
- The crust contains nearly all the mass

The lesson in all this is, we must be careful with our **geometric intuition** of high-dimensional spaces!

Paradoxes

Paradoxes

There are many more counter-intuitive examples in high-dimensional spaces, to list a few

Paradoxes

There are many more counter-intuitive examples in high-dimensional spaces, to list a few

- Most of the mass of a standard Gaussian is in the tail!

Paradoxes

There are many more counter-intuitive examples in high-dimensional spaces, to list a few

- Most of the mass of a standard Gaussian is in the tail!
- Rare events may not actually be that rare

Problems

Data get far apart fast

Data get far apart fast

Suppose we have data uniformly distributed on a 2-D grid, 10-D grid, 100-D grid or 1000-D grid.

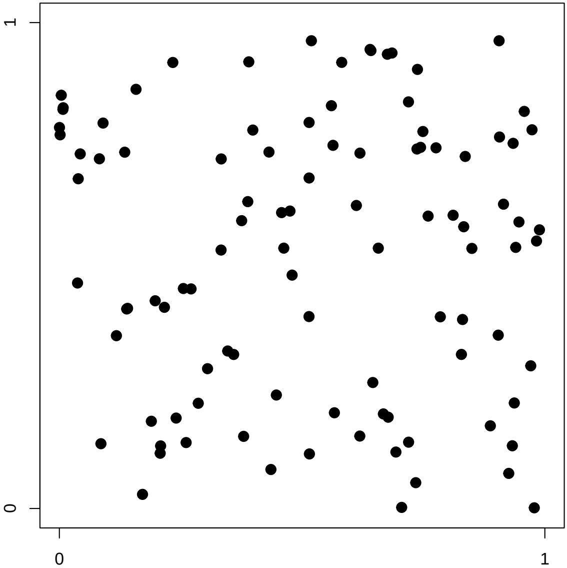
Data get far apart fast

Suppose we have data uniformly distributed on a 2-D grid, 10-D grid, 100-D grid or 1000-D grid.

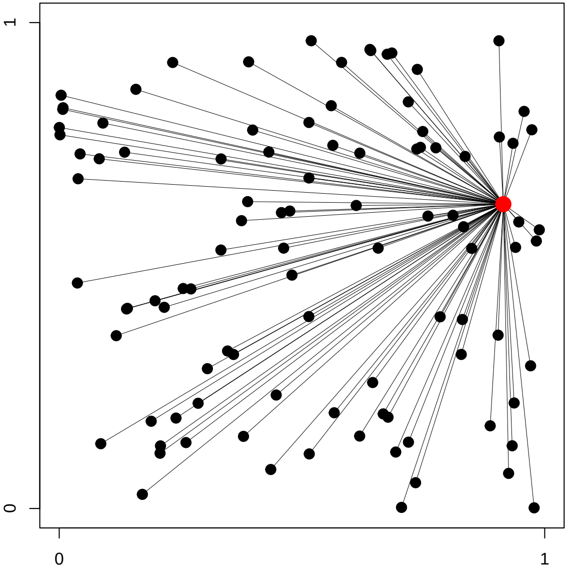
What's the average (euclidean) distance between points?

2D Case

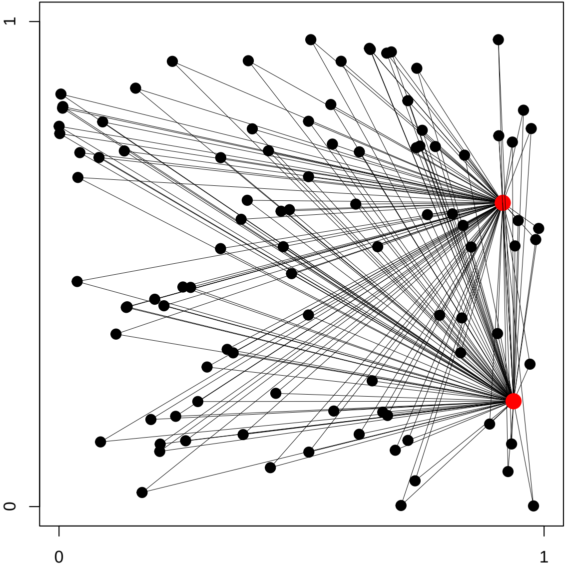
2D Case



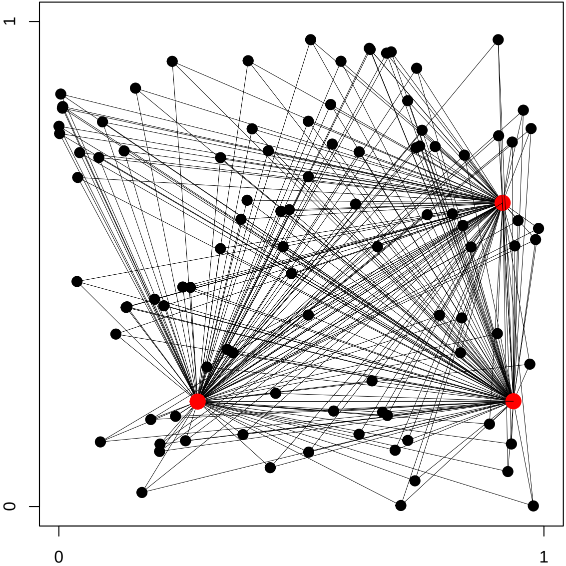
2D Case



2D Case

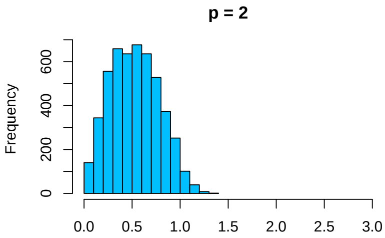


2D Case

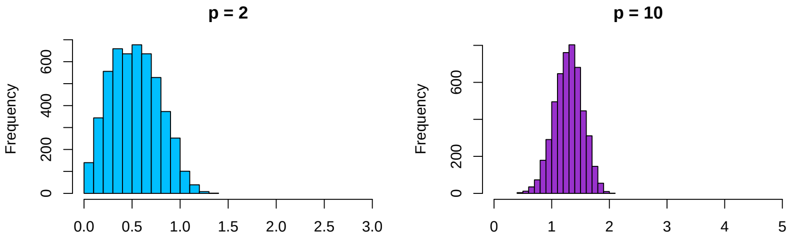


What's the average distance?

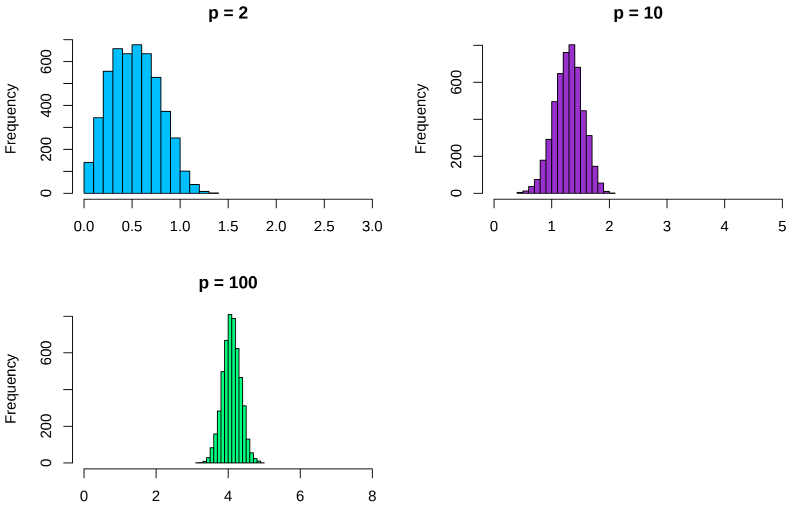
What's the average distance?



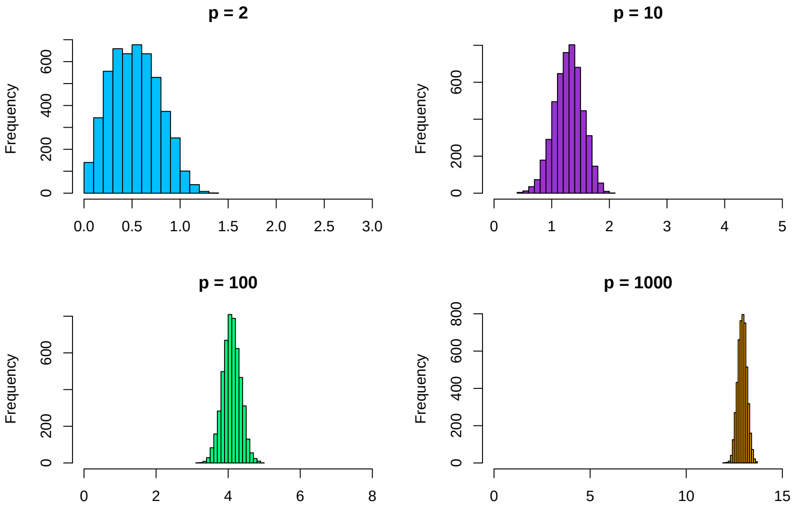
What's the average distance?



What's the average distance?

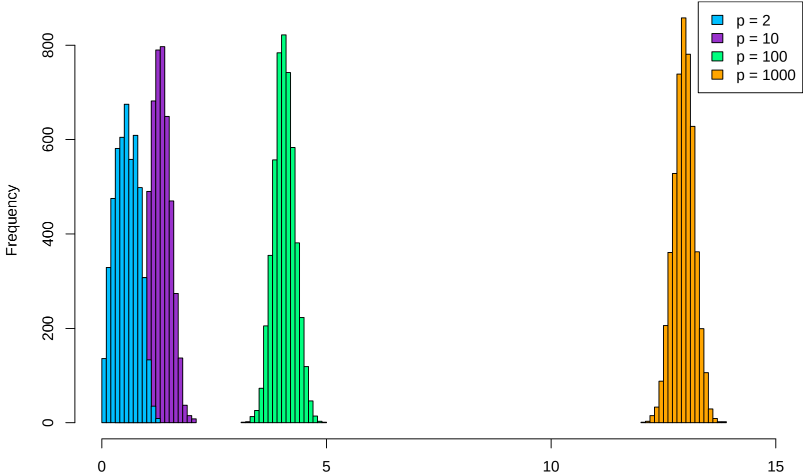


What's the average distance?



What's the average distance?

What's the average distance?



What does this mean?

What does this mean?

Distances between points

- The minimal distance between two points increases

What does this mean?

Distances between points

- The minimal distance between two points increases
- All points are a similar distance from the others

What does this mean?

Distances between points

- The minimal distance between two points increases
- All points are a similar distance from the others
- The notion of nearest point vanishes

Some more problems

Some more problems

High-dimensional spaces are immense:

- Vast, data points can be isolated in the immensity
- Small changes add up fast. Many small fluctuations in different directions can produce a large global fluctuation
- Computation suffers

A practical example

PCA

PCA

PCA relies on the estimation of the covariance matrix Σ.

PCA

PCA relies on the estimation of the covariance matrix Σ.

Let's suppose we have some some data generated iid from Normal$(0, I_p)$, where I_p is the $p \times p$ identity.

PCA

PCA relies on the estimation of the covariance matrix Σ.

Let's suppose we have some some data generated iid from Normal$(0, I_p)$, where I_p is the $p \times p$ identity.

How well is Σ estimated as p grows?

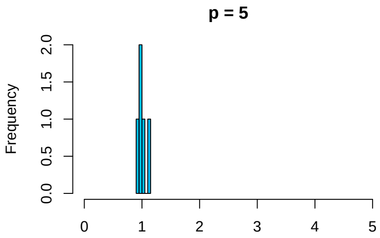
PCA

PCA

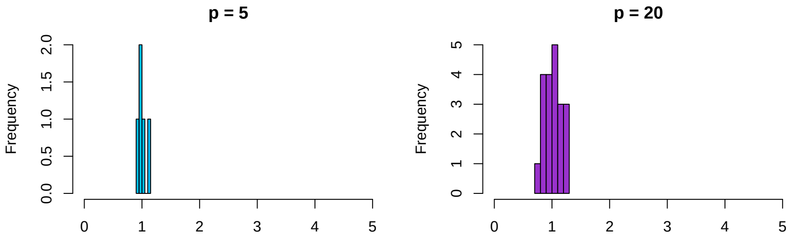
To see how well we're doing we're going to compare the eigenvalues of the estimated covariance matrices to that of the true covariance matrix (I_p), we let $n = 1000$ and let $p = 5, 20, 250, 500$.

PCA

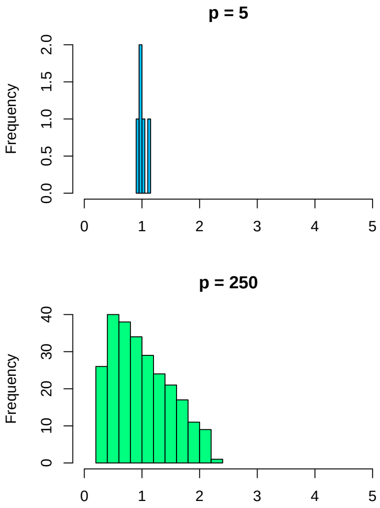
PCA



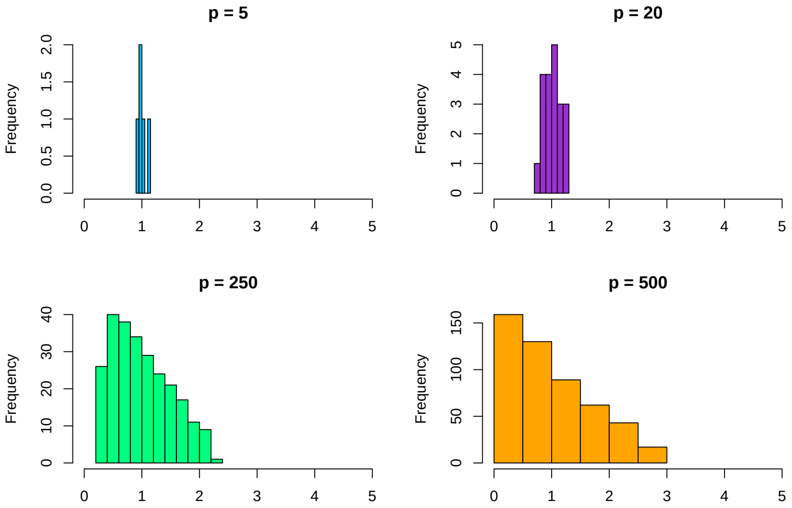
PCA


p = 5
Frequency
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
p = 20
Frequency
0
1
2
3
4
5
0
1
2
3
4
5

PCA


p = 5
Frequency
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
p = 20
Frequency
0
1
2
3
4
5
0
1
2
3
4
5
p = 250
Frequency
0
10
20
30
40
0
1
2
3
4
5

PCA


p = 5
p = 20
p = 250
p = 500
Frequency
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
10
20
30
40
50
100
150

PCA

PCA, main message

PCA

PCA, main message

The empirical covariance is a very poor approximation of the covariance I_p in this setting.

PCA

PCA, main message

The empirical covariance is a very poor approximation of the covariance I_p in this setting.

In turn, downstream use of the covariance matrix (e.g. in PCA) can lead to spurious results.

Silver lining

Silver lining

In many settings data is often much more low-dimensional and not uniformly spread!

Silver lining

In many settings data is often much more low-dimensional and not uniformly spread!

- Images have structures

Silver lining

In many settings data is often much more low-dimensional and not uniformly spread!

- Images have structures
- Biological systems are strongly regulated

Silver lining

In many settings data is often much more low-dimensional and not uniformly spread!

- Images have structures
- Biological systems are strongly regulated
- Consumption reflects some social structures

Silver lining

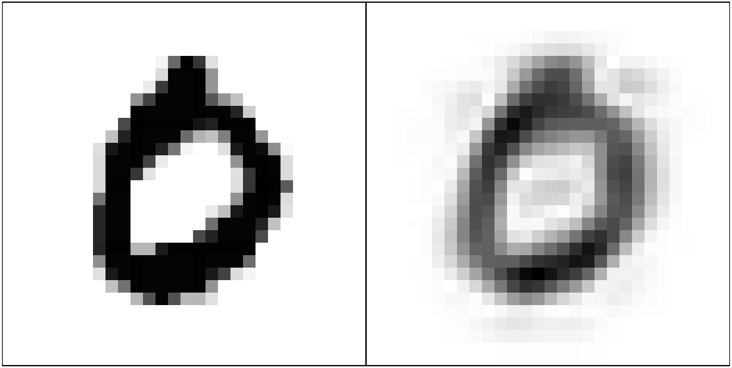
In many settings data is often much more low-dimensional and not uniformly spread!

- Images have structures
- Biological systems are strongly regulated
- Consumption reflects some social structures
- etc.

PCA on MNIST

PCA on MNIST

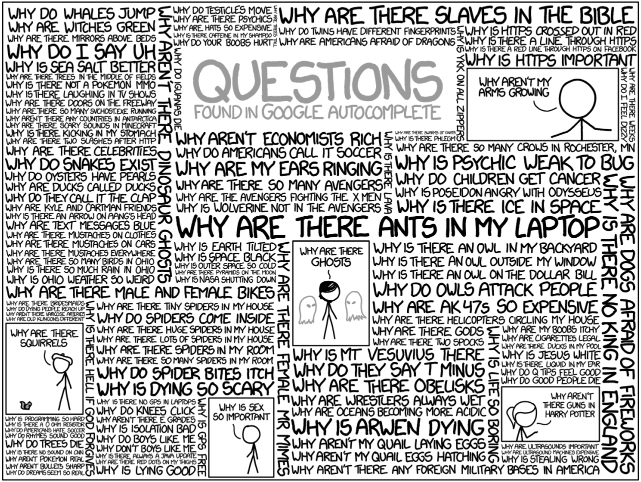
Reconstructing the MNIST digits from the first 25 PCs



Overall, we need to be careful when analyzing high-dimensional data

Overall, we need to be careful when analyzing high-dimensional data

Methods that can identify relevant structures or incorporate the knowledge that these spaces are often sparse generally outperform traditional methods.



QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47S SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE SQUIRRELS
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Reference I

[Gir21] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, Aug. 2021.